



ISSN: 2723-9535

Available online at [www.HighTechJournal.org](http://www.HighTechJournal.org)

# HighTech and Innovation Journal

Vol. 3, No. 1, March, 2022



## Dendrogram Analysis and Statistical Examination for Total Microbiological Mesophilic Aerobic Count of Municipal Water Distribution Network System

Mostafa Essam Eissa <sup>1\*</sup>, Engy Refaat Rashed <sup>2</sup>, Dalia Essam Eissa <sup>3</sup>

<sup>1</sup> *Independent Researcher, Pharmaceutical and Healthcare Research Facility, Cairo, Egypt.*

<sup>2</sup> *National Centre for Radiation Research and Technology, Cairo, Egypt.*

<sup>3</sup> *Royal Oldham Hospital, Rochdale Rd, Oldham OL1 2JH, United Kingdom.*

Received 23 November 2021; Revised 12 January 2022; Accepted 18 January 2022; Published 01 March 2022

### Abstract

The microbiological quality of water for human consumption is a critical safety aspect that should not be overlooked, especially when considering facilities for healthcare and the treatment of ill populations. Thus, the biological stability of water is crucial for the distribution network that delivers potable water to the final users for consumption and other human activities. The present work aimed to study a municipal distribution network system for city water within a healthcare facility. The implementation of the statistical analysis was conducted over long-term data collection, and the comparative study for the microbiological count of the water samples - from different points-of-use was assessed using the non-parametric analysis of the Kruskal-Wallis test. The comparative study involved a preliminary general one-way Analysis of Variance (ANOVA) followed by ad-hoc pairwise comparison. The statistical study involved a correlation matrix and a dendrogram to elucidate the level of association between different sections in the network. The ports C4 and C13 were at the trough in the microbiological count, in contrast to C13, which showed the highest level of the average microbial density. Despite a low to moderate level of correlation between the datasets of the water network, the tree diagram (dendrogram) analysis showed remarkable clustering. Use points could be grouped into three dense groups based on abrupt cuts in the similarity value. The study was useful in the analysis of the pattern and behavior of the microbial quality in a distribution water network in a specific area of the study. This work in turn would help in investigating the areas of improvement and defect spotting, in addition to assessing the biological stability of the water distribution system. The study could be extended to cover other different processed water networks, such as distilled, deionized, and purified water, as well as Water-For-Injection (WFI).

*Keywords:* Correlation Matrix; Dendrogram; Healthcare, Kruskal-Wallis Test; Microbiological Count; Municipal Water.

## 1. Introduction

In the world of the uninterrupted increase in the number of ill populations, immunocompromised patients, and defected healthy individuals, seeking appropriate control of human consumable products becomes a critical task, especially in the healthcare industry and settings [1]. One of the important components in everyday human activities is water [2]. There are several quality characteristics that must be considered in the monitoring and control of municipal water [3]. One of the most critical properties is the microbial count, or bioburden content, of the city water.

\* Corresponding author: [mostafaessameissa@yahoo.com](mailto:mostafaessameissa@yahoo.com)

 <http://dx.doi.org/10.28991/HIJ-2022-03-01-03>

➤ This is an open access article under the CC-BY license (<https://creativecommons.org/licenses/by/4.0/>).

© Authors retain all copyrights.

In this type of dynamic system, the biological stability of the water distribution network is essential to avoid any unexpected excursions in microbial quality [4]. The fewer fluctuations in the bioburden count (as frequency and magnitude), the less the risk of out-of-control (OOC) status might occur, the safer the water for human use and consumption [5]. Previous research has been carried out to monitor this approach using Statistical Process Control (SPC) and control or trending charts [6]. However, in the compound or complex system, it would be plausible to minimize the dimensionality of the variables through cluster analysis for the revealing of grouping in the characteristics in terms of microbial count [7]. In turn, this will highlight sections for improvement, defective regions, and acceptable parts.

Considering the importance of the homogeneity of the microbiological count in the water distribution network, the present study aimed to provide a statistical stability analysis of municipal water systems in a selected area of a healthcare facility. The work would cover a two-dimensional analysis of the pattern of the microbial count dispersion as a function of time and location. The distribution of data was analyzed and compared. Correlation and similarity studies could be executed to spot a unique grouping tendency in different segments of the water distribution network system.

## 2. Material and Methods

The study subject of this work was focused on the distribution network of the city water system in a selected healthcare facility. This network provides service for different partitions in the plant and each section possessed its own point-of-use port(s). The healthcare facility received its supply of municipal water lines from two sources. These stations were subject to regular monitoring for quality inspection characteristics [8]. An important monitored trait to be considered was the total microbiological aerobic, mesophilic viable count, or Total Viable Count (TVC).

The microbiological sampling was conducted aseptically over a 45-month period for the determination of the water plate count. The standard analysis technique for water samples collected in sterile bottles was conducted using an aseptic technique by a method provided by other researchers [9]. After incubation, the Heterotrophic Plate Count (HPC) was quantified as the number of Colony Forming Unit (CFU) per milliliter.

The reported data was collected chronologically in a column database for the plate count for each use point in the municipal water distribution system and processed using statistical software packages. The examination included general descriptive statistics, cumulative histograms, distribution pattern investigation, global variance analysis, multiple comparison tests, correlation matrix analysis, and tree diagram (Dendrogram). The programs included in the study embraced XLSTAT Premium version 2021.1.1 for Correlogram for correlation matrix drawing and P value plot [10]. This Excel-integrated software was also used to illustrate cumulative histograms for all dataset columns. GraphPad Prism for Windows version 6.01 [11] was used to create the descriptive statistics, graphical summary, overall bioburden comparison, and multiple pairwise comparisons in the table and figure. Finally, Minitab version 17.1.0 was used for cluster analysis using a tree diagram and to show a detailed tabulated similarity level [12].

## 3. Results and Discussion

The mean microbial count of city water for each point-of-use along with the overall value with Standard Error of the Mean (SEM) could be demonstrated in Figure 1. The maximum HPC was found at port C5, and the minimum average microbial levels were detected at C4 and C13. Figure 2 shows the distribution analysis of the microbial count as a cumulative histogram [13]. The most likely distribution fit for the discrete dataset was variable. However, the lognormal fit was the most common of the examined use points. Exceptions were found in C4, C9, and C13 with Weibull III, exponential, and Gamma II distributions, respectively. In the same line, Figure 3 illustrates the pattern of data using a probability–probability or percent–percent (P-P) plot. All dispersions were akin to each other, suggesting close spreading of the datasets [14]. All the results were far from the Gaussian distribution's usual shape, and this finding was expected based on the previously found dispersion pattern in other studies of the microbial count distribution in the water samples.

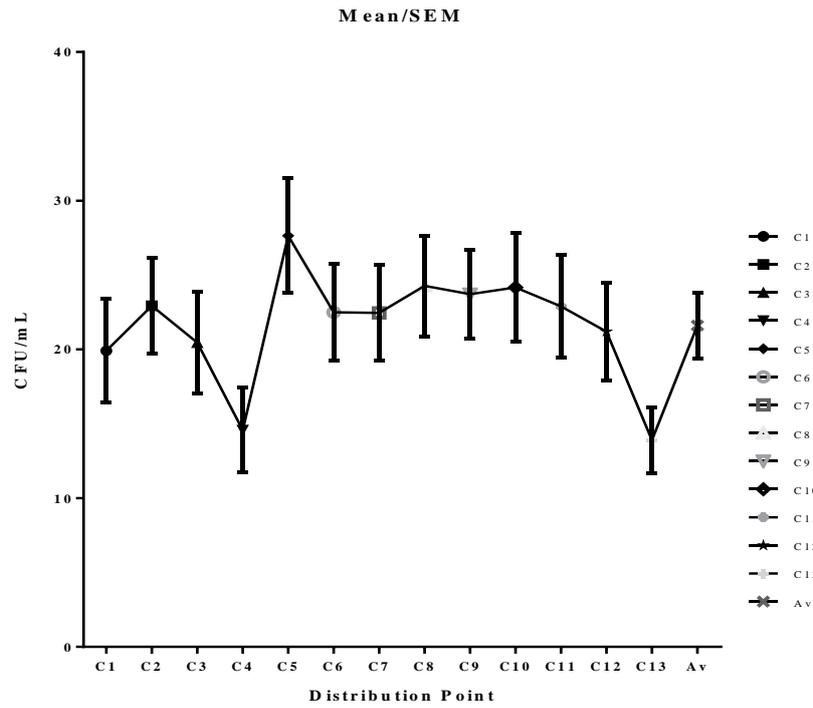


Figure 1. Heterotrophic Plate Count (HPC) as a mean CFU/mL ± Standard Error of the Mean (SEM) of 13 distribution points (C) with the average (Av) is shown for the whole Municipal system

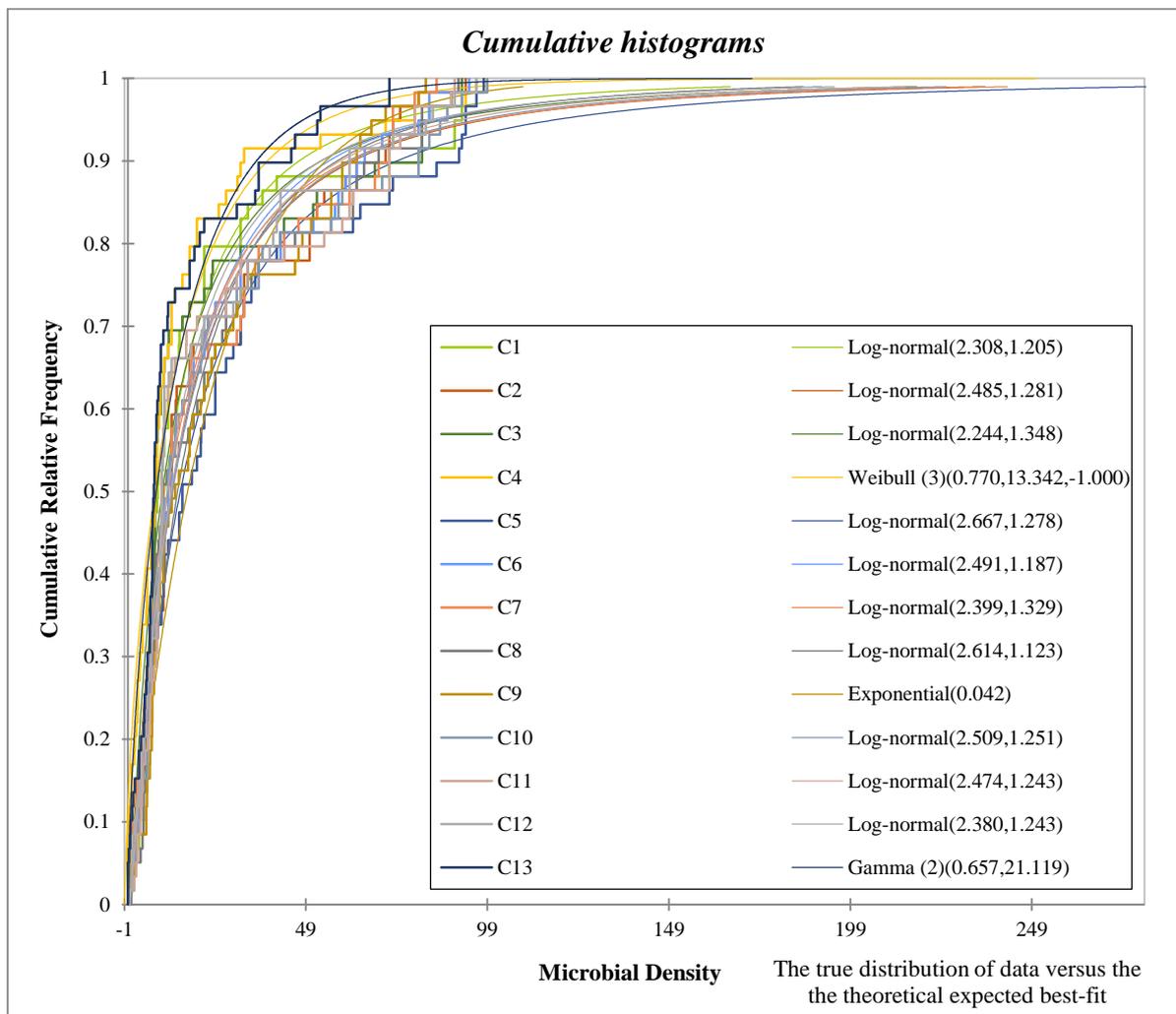
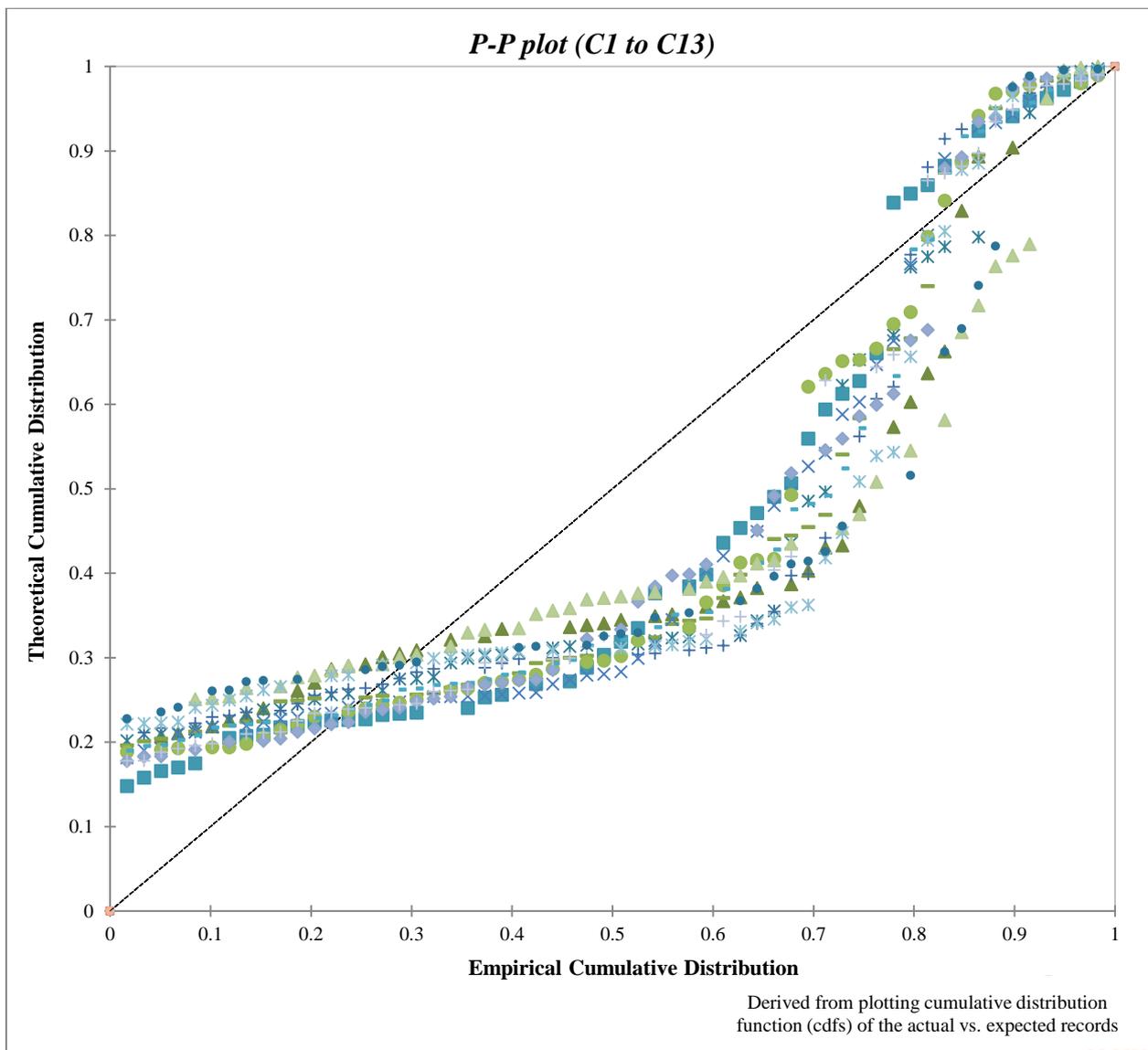


Figure 2. Cumulative histogram of the distribution points of municipal water network in the healthcare facility

Multiple comparison analysis was investigated between the microbial count trend of all working lines of the city water distribution system within the healthcare facility. The approximate  $P \leq 0.05$  was found to be 0.0187 using Analysis of Variance (ANOVA). The general (global) effect is pinpointed as a significant variation in microbial quality between different sections of the water distribution network in the plant due to significant variation between medians. Nevertheless, a multiple comparison (post hoc) test using a non-parametric test Kruskal-Wallis for pairwise analysis yielded a non-significant difference (Figure 4). Thus, the conclusion of significant ANOVA with non-significant multiple pairwise comparisons was that the p-value computed by the ANOVA was lower than the alpha ( $\alpha$ ) significance level (e.g. 0.05) [15]. All the p-values computed by the pairwise multiple comparisons test were higher than the  $\alpha$  significance level.

Some reasons for why post-hoc test might appear not significant while the overall effect was significant. A conservative multiple comparisons test. A weakly significant global effect (p-value of the ANOVA table was very close to the significance level) was not the case in the present study. Hence, this reason was excluded. Another reason that should be investigated was the lack of statistical power. For instance, when treatments had small sizes. When multiple comparisons tests were not statistically powerful, it would be less likely to detect significant differences. However, this was not the case herein also as the number of values per group was reasonably high. The more conservative the test, the more likely rejection would be found significantly different between groups that in reality were meaningful. In addition, a high number of factor levels can also be an explanation as in the current situation of Table 1. The more pairwise comparisons would be found in hand, the more p-values will get penalized in order to decrease the risk of rejecting null hypotheses while they are true [15]. Thus, there is a reasonable assumption for considering a significant variation in the microbiological water quality between different sections in the distribution network.



**Figure 3. Probability–Probability (P-P) plot of the actual versus the expected data dispersion of the point-of-use ports spreading across water distribution system in the plant**

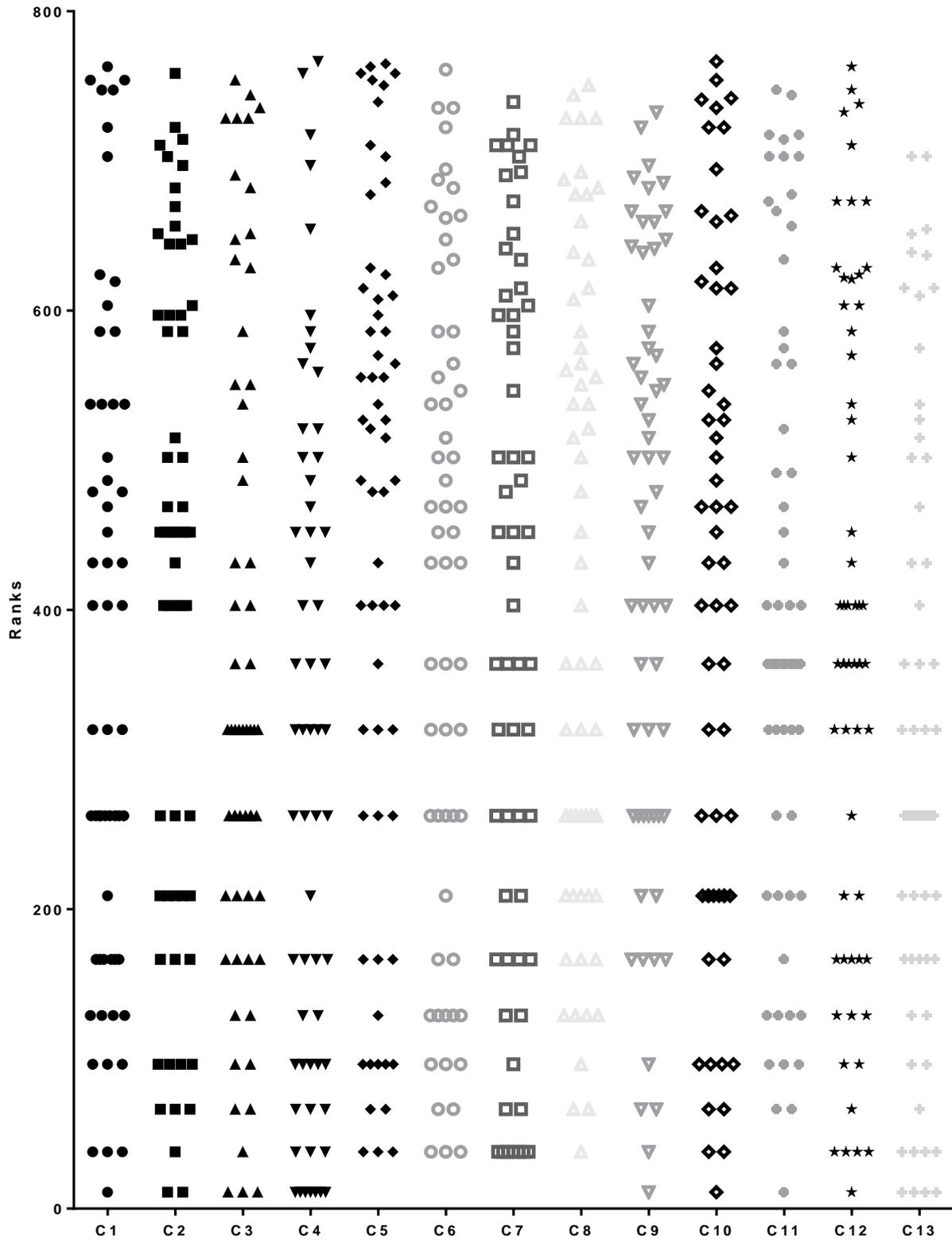


Figure 4. Kruskal-Wallis test ( $p < 0.05$ ) showing ranking distribution for all use points of water distribution system

The correlation matrix between different segments of the municipal water distribution system is shown in Figure 5 [16]. It could be noted that a low correlation level existed between the point-of-use ports with moderate records that were observed at the best estimates.

**Table 1. Kruskal-Wallis test for 13 treatment (columns) of 767 total values with 78 multiple pairwise comparisons per family**

Test details *	Mean rank I	Mean rank II	Mean rank diff. †	Test details *	Mean rank I	Mean rank II	Mean rank diff. †
C1 vs. C2	363.6	405.4	-41.72	C4 vs. C11	302.0	398.9	-96.86
C1 vs. C3	363.6	357.6	6.017	C4 vs. C12	302.0	382.7	-80.70
C1 vs. C4	363.6	302.0	61.64	C4 vs. C13	302.0	311.2	-9.23
C1 vs. C5	363.6	437.2	-73.51	C5 vs. C6	437.2	398.9	38.24
C1 vs. C6	363.6	398.9	-35.27	C5 vs. C7	437.2	389.3	47.83
C1 vs. C7	363.6	389.3	-25.68	C5 vs. C8	437.2	410.8	26.31
C1 vs. C8	363.6	410.8	-47.19	C5 vs. C9	437.2	430.5	6.644
C1 vs. C9	363.6	430.5	-66.86	C5 vs. C10	437.2	403.8	33.35
C1 vs. C10	363.6	403.8	-40.16	C5 vs. C11	437.2	398.9	38.29
C1 vs. C11	363.6	398.9	-35.22	C5 vs. C12	437.2	382.7	54.44
C1 vs. C12	363.6	382.7	-19.07	C5 vs. C13	437.2	311.2	125.90
C1 vs. C13	363.6	311.2	52.41	C6 vs. C7	398.9	389.3	9.59
C2 vs. C3	405.4	357.6	47.74	C6 vs. C8	398.9	410.8	-11.92
C2 vs. C4	405.4	302.0	103.40	C6 vs. C9	398.9	430.5	-31.59
C2 vs. C5	405.4	437.2	-31.79	C6 vs. C10	398.9	403.8	-4.8900
C2 vs. C6	405.4	398.9	6.449	C6 vs. C11	398.9	398.9	0.05
C2 vs. C7	405.4	389.3	16.04	C6 vs. C12	398.9	382.7	16.20
C2 vs. C8	405.4	410.8	-5.48	C6 vs. C13	398.9	311.2	87.68
C2 vs. C9	405.4	430.5	-25.14	C7 vs. C8	389.3	410.8	-21.52
C2 vs. C10	405.4	403.8	1.56	C7 vs. C9	389.3	430.5	-41.19
C2 vs. C11	405.4	398.9	6.50	C7 vs. C10	389.3	403.8	-14.48
C2 vs. C12	405.4	382.7	22.65	C7 vs. C11	389.3	398.9	-9.54
C2 vs. C13	405.4	311.2	94.13	C7 vs. C12	389.3	382.7	6.61
C3 vs. C4	357.6	302.0	55.62	C7 vs. C13	389.3	311.2	78.08
C3 vs. C5	357.6	437.2	-79.53	C8 vs. C9	410.8	430.5	-19.67
C3 vs. C6	357.6	398.9	-41.29	C8 vs. C10	410.8	403.8	7.03
C3 vs. C7	357.6	389.3	-31.69	C8 vs. C11	410.8	398.9	11.97
C3 vs. C8	357.6	410.8	-53.21	C8 vs. C12	410.8	382.7	28.13
C3 vs. C9	357.6	430.5	-72.88	C8 vs. C13	410.8	311.2	99.60
C3 vs. C10	357.6	403.8	-46.18	C9 vs. C10	430.5	403.8	26.70
C3 vs. C11	357.6	398.9	-41.24	C9 vs. C11	430.5	398.9	31.64
C3 vs. C12	357.6	382.7	-25.08	C9 vs. C12	430.5	382.7	47.80
C3 vs. C13	357.6	311.2	46.39	C9 vs. C13	430.5	311.2	119.30
C4 vs. C5	302.0	437.2	-135.1	C10 vs. C11	403.8	398.9	4.94
C4 vs. C6	302.0	398.9	-96.91	C10 vs. C12	403.8	382.7	21.09
C4 vs. C7	302.0	389.3	-87.31	C10 vs. C13	403.8	311.2	92.57
C4 vs. C8	302.0	410.8	-108.80	C11 vs. C12	398.9	382.7	16.15
C4 vs. C9	302.0	430.5	-128.50	C11 vs. C13	398.9	311.2	87.63
C4 vs. C10	302.0	403.8	-101.80	C12 vs. C13	382.7	311.2	71.47

\* Test statistic = 24.27;

† No significant difference at  $\alpha = 0.05$

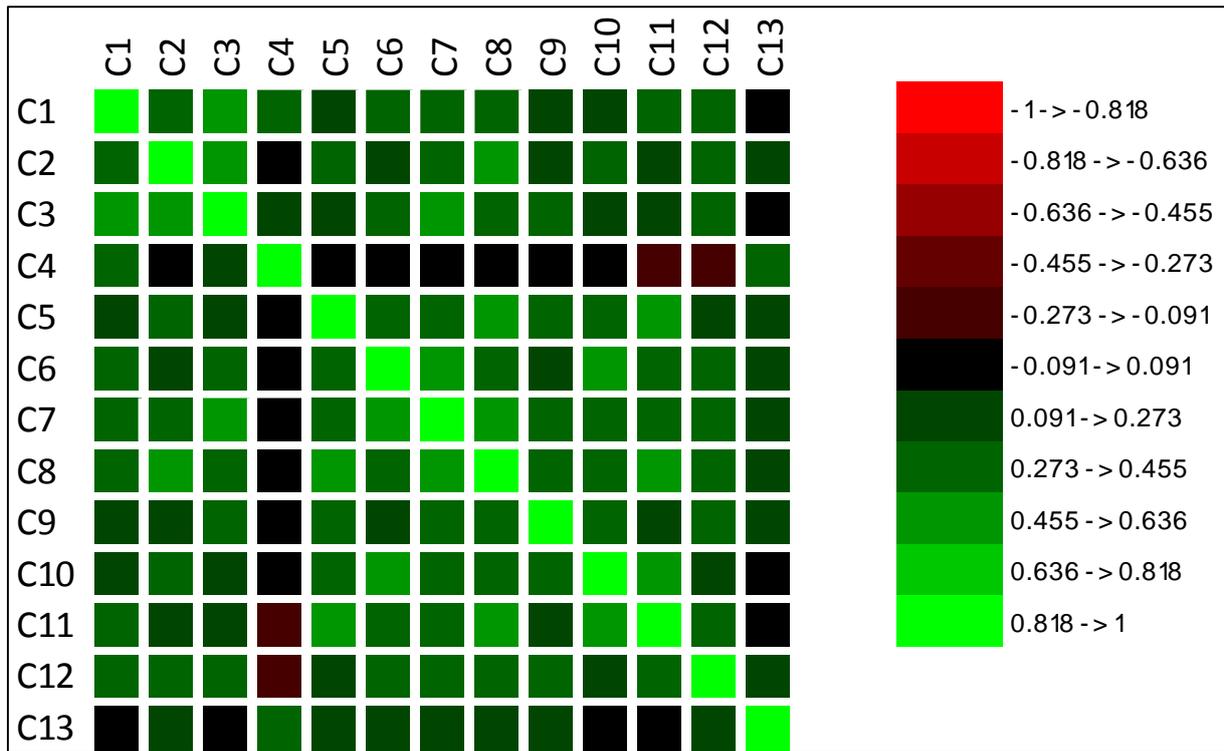


Figure 5. Spearman correlation matrix showing the degree of association of the total microbiological count between studied group of the point-of-use in a healthcare facility

Table 2 showed the formation of clusters for dendrogram creation at each step and determined the similarity (or distance) levels of the clusters formed. The pattern of how similarity or distance values change from step to step could aid in the selection of the final clustering for the database [17]. The step where the values changed abruptly might be identified as an acceptable point to define the final grouping. The decision about final grouping was also called “cutting the dendrogram”. Cutting the dendrogram was akin to drawing a line across the dendrogram to specify the final clustering [17]. In this essence, the inflection between step 10 (8 observations) and 11 (12 observations) could be observed. The dendrogram (tree diagram) was used to display the groups formed by aggregation of variables at each increment and showed their similarity levels. In Figure 6, the similarity levels – which could be displayed as distance level as well - were quantified along the y-axis through measuring the corresponding horizontal line at each step and the various variables (distribution network sampling points) were listed along the x-axis. Accordingly, three cutting edge clusters could be identified viz. 1(4), 5(8) and 13(1).

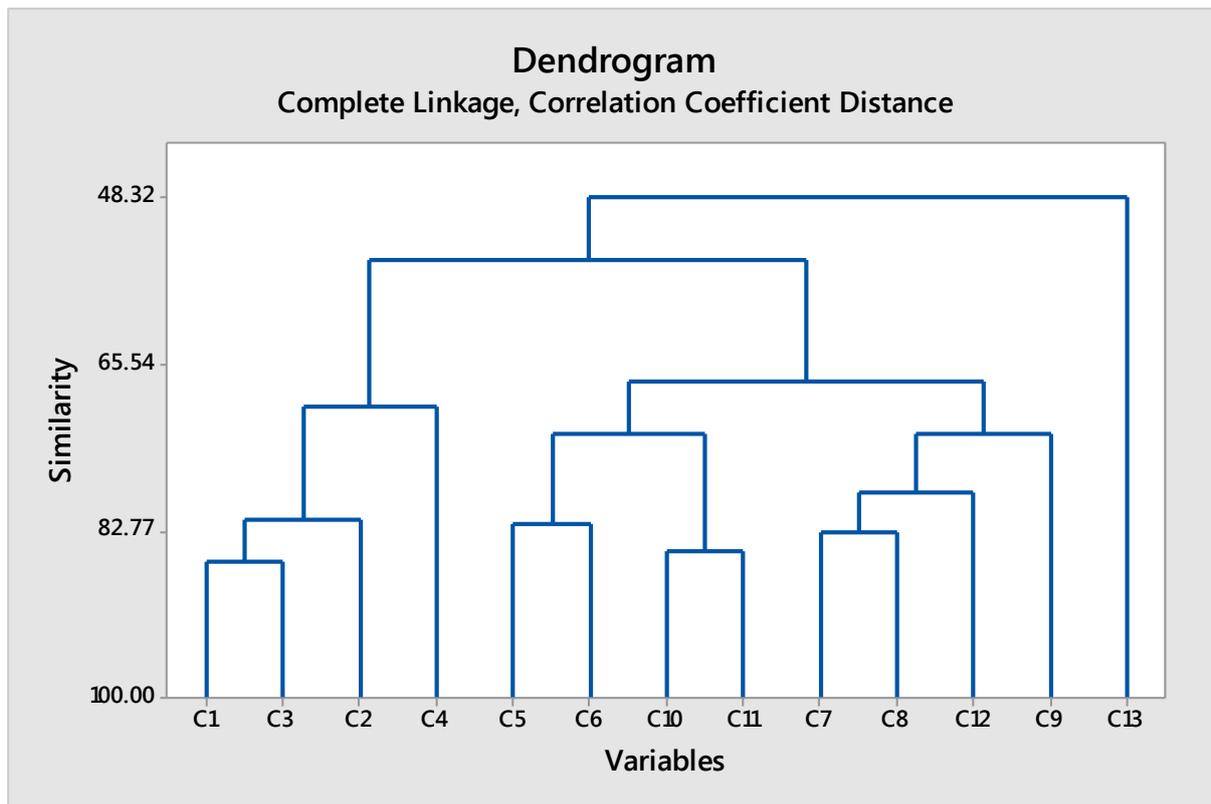
Table 2. Correlation Coefficient distance, complete linkage and amalgamation Steps

Cluster Analysis of Variables: C1→C13						
Step	Number of clusters	Similarity level	Distance level	Clusters Joined *	New cluster †	Number € of obs.in new cluster
1	12	85.8623	0.28275	1→3	1	2
2	11	84.7412	0.30518	10→11	10	2
3	10	82.9348	0.34130	7→8	7	2
4	9	81.9826	0.36035	5→6	5	2
5	8	81.5184	0.36963	1→2	1	3
6	7	78.8315	0.42337	7→12	7	3
7	6	72.7648	0.54470	7→9	7	4
8	5	72.7235	0.54553	5→10	5	4
9	4	69.8196	0.60361	1→4	1	4
10	3	67.3546	0.65291	5→7	5	8
11	2	54.7786	0.90443	1→5	1	12
12	1	48.3157	1.03369	10→13	1	13

\* The group of pair sets that were linked to yield a different cluster at each level in the amalgamation operation.

† The unique number of the developed cluster that was created at each level in the process of amalgamation.

€ The number of records in each newly created cluster at each incremental level in the process of amalgamation.



**Figure 6. Dendrogram showing similarity level of the microbiological count between the examined distribution points of the water network in a healthcare plant.**

While the present analysis is limited by the TVC only in this work, the study could be extended in the future to cover other quality aspects of water such as Total Organic Carbon (TOC) and conductivity, in addition to the traceability of specific objectionable and pathogenic microorganisms that are a potential risk to human and other living organisms' health in water. The advantage of the combination of these statistical techniques with the new technologies in microbial enumeration and detection should not be underestimated.

#### 4. Conclusion

Monitoring the biological stability of the water distribution system is crucial for the safety of human consumption and other associated activities. One of the important quality criteria of the distribution network is the control and monitoring of HPC. Statistical analysis of a well-established database might reveal useful information for reporting the quality issues, patterns, and trends that would help and support decision-making in continuous improvement projects. A single type of analysis might not reveal outcomes that could be revealed by another one, such as global ANOVA against pairwise multiple comparisons and correlation matrix against dendrogram. The overall comparison revealed a significant variation between different parts of the distribution system, yet this was not evident in the stepwise comparative analysis. In the same line, the correlation matrix did not yield sufficiently interesting associations in the present situation between different water sampling ports that could demonstrate a pattern. In the existing situation, a long-term study of municipal distribution networks showed a clustering tendency in the system segments related to the bioburden level, identifying three main groups. While there could be found two clusters of four and eight points-of-use, the last group consisted of only one line that stood at the end of the system, segregated from the adjacent network groups.

#### 5. Declarations

##### 5.1. Author Contributions

Conceptualization, D.E.E. and E.R.R.; methodology, M.E.E.; software, M.E.E.; validation, E.R.R., D.E.E. and M.E.E.; formal analysis, M.E.E.; investigation, E.R.R.; resources, D.E.E.; data curation, M.E.E.; writing—original draft preparation, M.E.E.; writing—review and editing, E.R.R.; visualization, E.R.R.; supervision, E.R.R.; project administration, D.E.E.; funding acquisition, D.E.E. All authors have read and agreed to the published version of the manuscript.

##### 5.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author

### 5.3. Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

### 5.4. Institutional Review Board Statement

Not Applicable

### 5.5. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## 6. References

- [1] World Health Organization. (2021). Health21: the health for all policy framework for the WHO European Region. Available online: [https://www.euro.who.int/\\_\\_data/assets/pdf\\_file/0010/98398/wa540ga199heeng.pdf](https://www.euro.who.int/__data/assets/pdf_file/0010/98398/wa540ga199heeng.pdf) (accessed on December 2021).
- [2] Water Science School. (2019). The Water in You: Water and the Human Body | U.S. Geological Survey. Available online: <https://www.usgs.gov/special-topics/water-science-school/science/water-you-water-and-human-body> (accessed on December 2021).
- [3] Hanaor, D., & Sorrell, C. (2013). Sand Supported Mixed-Phase TiO<sub>2</sub> Photocatalysts for Water Decontamination Applications. *Advanced Engineering Materials*, 16(2), 248-254. doi:10.1002/adem.201300259
- [4] Nescerecka, A., Rubulis, J., Vital, M., Juhna, T., & Hammes, F. (2014). Biological Instability in a Chlorinated Drinking Water Distribution Network. *Plos ONE*, 9(5), e96354. doi:10.1371/journal.pone.0096354
- [5] Bartram, J., Cotruvo, J. A., Exner, M., Fricker, C., & Glasmacher, A. (Eds.). (2003). *Heterotrophic plate counts and drinking-water safety*. IWA Publishing, London, United Kingdom.
- [6] Essam Eissa, M. (2017). Determination of the Microbiological Quality of Feed City Water to Pharmaceutical Facility: Distribution Study and Statistical Analysis. *Athens Journal of Sciences*, 4(2), 143-160. doi:10.30958/ajs.4-2-4
- [7] Paliy, O., & Shankar, V. (2016). Application of multivariate statistical techniques in microbial ecology. *Molecular Ecology*, 25(5), 1032-1057. doi:10.1111/mec.13536
- [8] Eissa, M. E. A. (2018). Microbiological quality of purified water assessment using two different trending approaches: A case study. *Sumerianz Journal of Scientific Research*, 1(3), 75-79.
- [9] Jyothy, S., Chandran, A., Abbas, A., & Hatha, A. M. (2012). Survival of *Escherichia coli*, *Aeromonas hydrophila* and *Pseudomonas fluorescens* in Bottled Water and Biofilms. *Journal of Nature Science and Sustainable Technology*, 6(1), 25.
- [10] XLSTAT 2020. (2020). XLSTAT Support Center. Available online: [https://help.xlstat.com/s/article/download-the-xlstat-help-documentation?language=en\\_US](https://help.xlstat.com/s/article/download-the-xlstat-help-documentation?language=en_US) (accessed on December 2021).
- [11] GraphPad Prism V6.01. (2014). GraphPad Prism User Guide. Available online: <https://cdn.graphpad.com/docs/prism/6/Prism-6-User-Guide.pdf> (accessed on December 2021).
- [12] Brook, Q. (2006). *Six sigma and Minitab*. QSB Consulting, Hampshire, United Kingdom.
- [13] A complete statistical add-in for Microsoft Excel. (2021). Histograms. Available online: <https://www.xlstat.com/en/solutions/features/histograms> (accessed on December 2021).
- [14] De Giglio, O., Diella, G., Lopuzzo, M., Triggiano, F., Calia, C., & Pousis, C. et al. (2021). Management of Microbiological Contamination of the Water Network of a Newly Built Hospital Pavilion. *Pathogens*, 10(1), 75. doi: 10.3390/pathogens10010075
- [15] XLSTAT (2020), How to Interpret Contradictory Results between Anova and Multiple Pairwise Comparisons? (2020). XLSTAT Support Center. Available online: [https://help.xlstat.com/s/article/how-to-interpret-contradictory-results-between-anova-and-multiple-pairwise-comparisons?language=en\\_US](https://help.xlstat.com/s/article/how-to-interpret-contradictory-results-between-anova-and-multiple-pairwise-comparisons?language=en_US) (accessed on December 2021).
- [16] XLSTAT (2020), Spearman Correlation Coefficient in Excel Tutorial. (2020). XLSTAT Support Center. Available online: [https://help.xlstat.com/s/article/spearman-correlation-coefficient-in-excel-tutorial?language=en\\_US](https://help.xlstat.com/s/article/spearman-correlation-coefficient-in-excel-tutorial?language=en_US) (accessed on December 2021).
- [17] Minitab®18 Support. (2019). Interpret all statistics and graphs for Cluster Variables - Minitab. Available online: <https://support.minitab.com/en-us/minitab/18/help-and-how-to/modeling-statistics/multivariate/how-to/cluster-variables/interpret-the-results/all-statistics-and-graphs/> (accessed on December 2021).